United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research Division

SRB Research Report
Number SRB 92-11

December 1992

EVALUATING THE ADDITION OF WEATHER DATA TO SURVEY DATA TO FORECAST SOYBEAN YIELDS

M. Denice McCormick
Thomas R. Birkett

EVALUATING THE ADDITION OF WEATHER DATA TO SURVEY DATA TO FORECAST SOYBEAN YIELDS, by M. Denice McCormick, Sampling and Estimation Research Section, Research Division, National Agricultural Statistics Service, Fairfax, Virginia 22030, and Thomas R. Birkett, Yield and Labor Section, Statistical Methods Branch, Estimates Division, National Agricultural Statistics Service, U.S. Department of Agriculture, Washington, D.C. 20250-2000, Research Report No. SRB 92-11, December, 1992.

ABSTRACT

The National Agricultural Statistics Service conducts surveys, during the growing season, to collect plant counts and measurements for forecasting yields of major agricultural crops such as corn, cotton, soybeans and wheat. Plot level data are aggregated to the regional (multi-state) level and used to build regression forecast models. In order to improve the accuracy of August 1 and September 1 yield forecasts, a cumulative precipitation term, indicating total precipitation from April 1 until the forecast date, was added to the models at a regional six-state level. This additional data was believed to contain information which would be helpful in crop yield forecasting. The analysis indicated that the accuracy of the regional soybean yield forecasts was not improved using this precipitation term. Additional analysis is recommended to evaluate the impact of different precipitation terms on soybean yield forecasts at the regional and state levels, and to evaluate the use of precipitation data in forecasting crop yields for corn, cotton and wheat.

KEY WORDS

Precipitation, regression, model evaluation.

This paper was prepared for limited distribution to the research community outside the U.S. Department of Agriculture. The views expressed herein are not necessarily those of NASS or USDA.

ACKNOWLEDGEMENTS

The authors wish to thank Fred Warren, retired Mathematical Statistician, National Agricultural Statistics Service, for providing insight during the preliminary stages of this study, and to Lee Brown, Carol House, Bill Iwig and Phil Kott for their assistance and commitment to this research project.

TABLE OF CONTENTS

Page

TABLES

SUMMARY

The National Agricultural Statistics Service (NASS) conducts Objective Yield (OY) surveys, during the growing season, to collect plant counts and measurements used to forecast yields of major agricultural crops such as wheat, corn, cotton, and soybeans. The collected plot data are aggregated to the regional (multi-state) level and then used to build regression forecasting models under a structure introduced by Birkett (1990). Regional values of survey variables are regressed against the final regional yield published by the Agricultural Statistics Board. State yields are then forecasted using an allocation process, under the condition that the State yields must weight to the regional yield. A different model structure based on plot level regression models, was used previous to 1990 and continues to provide alternative yield forecasts. Although the new approach provides more accurate forecasts, there is still potential for improvement.

This analysis evaluates the addition of a cumulative precipitation term to the regional soybean models covering Illinois, Indiana, Iowa, Minnesota, Missouri, and Ohio for the August 1 and September 1 forecasts. The precipitation term is total precipitation from April 1 until the forecast date aggregated to the regional level. Data from 1980 through 1991 were used for this analysis. Five different models were evaluated for each month using different linear, quadratic, and interaction terms of the precipitation and survey variables. Evaluation criteria include a comparison of the length of the prediction interval, the adjusted coefficient of determination (R_a^2), the average absolute relative difference (ARD) between the prediction and Board yield, and the number of years the ARD is less than 5%. The prediction intervals were calculated for 1988, 1981, and 1990, which are years that represent the occurrence of the minimum, median, and maximum six state regional soybeans yields, respectively, for the years covered in this analysis.

For August, the analysis indicated that the best forecast model at the regional level is the simple linear regression model using the number of lateral branches per eighteen square feet. This is the regional model currently used by NASS for August 1 yield forecasts. This model consistently produced the smallest prediction interval all under 3 bushels per acre, for the three years examined. Adding the precipitation term to this model increased the length of the prediction intervals for all three years, but did produce an equivalent R_a^2 and lower ARD values. It was also determined that the predictor variables, lateral branches and precipitation, are not strictly independent of one another. However, this collinearity problem (lack of strict independence) is not troublesome. Both variables are positively correlated to Board yield.

For September, the analysis indicated that a quadratic model using linear and squared terms of the survey variable is the best forecast model for soybeans at the regional level. The survey variable used in September is the number of pods per eighteen square feet. This model was used on an experimental basis by NASS for the September 1, 1991 soybean yield forecast. The standard September model is the simple linear model.

The experimental model consistently produced the smallest prediction interval. With the outlier year 1980 excluded from the analysis, the length of the prediction interval was approximately 1.5 bushels per acre. Adding the precipitation term to this quadratic model provided nearly identical R_a^2 and ARD values as the quadratic model, but still increased the prediction intervals for all three years examined. As in August, the predictor variables, pods and precipitation, were found not to be strictly independent of each other. The correlation between Board yield and both predictor variables increased for September, and were both found to be significant.

In conclusion, there is no evidence that a change from the univariate survey data model, defined as a simple linear regression model using the number of lateral branches per eighteen square feet, is warranted for the August forecast period. For the September forecast period, the quadratic model, using pods and pods^2, shows definite improvement in all evaluation criteria over the univariate model. Adding the precipitation term investigated for this study to the quadratic model shows no gain in forecast accuracy at the regional level.

Based on this analysis, the following recommendations are made.

1. Investigate other precipitation time frame terms such as monthly total accumulated precipitation for May, June, July and August as well as row spacing and planting dates.

2. Analyze other crops such as corn, cotton and wheat to determine if weather data can improve forecast accuracy at the regional and state levels.

3. Continue research to analyze the feasibility of a real time weather data system whereby preliminary precipitation data are used in conjunction with historical weather data to make current yield forecasts.

EVALUATING THE ADDITION OF WEATHER DATA TO SURVEY DATA TO FORECAST SOYBEAN YIELDS

M. Denice McCormick

and

Thomas R. Birkett

INTRODUCTION

In 1990, the National Agricultural Statistics Service (NASS) introduced new Objective Yield (OY) models to forecast yield for corn and soybeans on the regional and state levels in a plan to phase out the older, less accurate operational models (Birkett 1990). The Objective Yield Survey collects data from randomly selected sample plots in randomly selected fields. The old regression models predicted the components of yield such as number of pods per plant and weight per pod at the plot level based on five years of previous data. Plot level data were then aggregated to the state level. The new models are also regression models, and have initially been developed to predict yield directly rather than the components of yield using survey data aggregated to the regional level. Regions are constructed from states in the Objective Yield program. A longer period of years in the historic data set must be used since only one data point is used to represent each year. State level forecasts are modeled subject to the linear constraint that they weight to the regional forecast. Yield component models under the new model structure are currently being investigated. Analysis has indicated that these new models are more accurate than the old models. Nevertheless, there is potential for improvement, particularly in the area of making early season forecasts and forecasting yield for outlier years.

Discussions in March 1991 with Professor James Beuerlein of the Agronomy Department of the Ohio State University provided information that suggested the use of weather data along with survey data to improve early season forecasts. Additionally, row space measurements and planting dates were also indicated as variable candidates.

This research effort evaluates the addition of precipitation data to the new Objective Yield model for soybeans, in order to improve the precision of early season (August 1 and September 1) yield forecasts. This investigation considers data for twelve years, 1980 to 1991, for a region of six states that participate in the annual Objective Yield Survey: Illinois, Indiana, Iowa, Minnesota, Missouri and Ohio. The performance of the models is examined in this report for August and September, the early season forecast periods.

Attempts have been made previously to include weather data in Objective Yield models. Sanderson (1942) used crop condition reports and weather data to forecast the yield per acre of wheat and found gains could be made in forecast accuracy, especially in late season models. House (1977) recommended that weather variables be incorporated into a within-year growth model to forecast corn yields. Sebaugh (1981) conducted a number of investigations in this area. In one study, she included weather data in the analysis of the performance of Climatic and Environmental Assessment Services (CEAS) and Thompson models (1981) in forecasting spring wheat yields. She also analysed the ability of Kestle's "Straw Man" model to forecast corn and soybean yields using weather data (1981). Later, Sebaugh and Cotter investigated models containing weather data that forecasted soybeans (1983). Others, such as Maas (1982), Sebaugh (1983), and Warren (1990) constructed weather related indices to include in yield forecast models. To date, most of the previous research has not provided significant improvement in crop yield forecasting. Possible problem areas included unreliability and inadequate coverage of weather data over the forecast area and improper data aggregation or model structure.

One concern in the research and application of forecast techniques is the difficulty of obtaining timely and reliable weather data. In this study, the data were obtained from the National Oceanic and Atmospheric Administration (NOAA), which collects historical weather data through the National Weather Service system. NOAA collects, edits, and then stores the data for public use. Unfortunately, NOAA cannot provide this information to the public in less than three to six months from the date of collection. Preliminary data are collected through a regional reporting system. Not all stations report on a daily basis. This is because the extended network of weather data reporters includes a mixture of both human non-paid volunteers and automated systems. Also, since the majority of stations are operated by volunteers, there are occasions when reporting stations come in and out of the system. At the end of each month, a summary is sent from each reporting station to the regional National Weather Service office. A regional summary is then sent to NOAA where a final data set is compiled. The process to build a final, relatively reliable data set takes a minimum of six months.

A second concern is the coverage of weather data. The National Weather Service has developed a system of climatic divisions which frequently does not coincide with the system of mutually exclusive Agricultural Statistics Districts (ASD) established in each State by NASS. Reporting stations report measurements taken for specific locations, which NOAA claims to provide adequate coverage based on their climatic divisions. It is unclear whether coverage is still adequate for all ASDs. When working with the data, an assumption is made that each ASD has a representative number of reporting stations for each time period. Approximately nine stations per ASD is assumed to be an acceptable level of coverage for this study. To examine whether this assumption is met, the coverage for Ohio was studied. Ohio has nine ASDs, averages nine to ten counties per ASD, and is assumed to be fairly representative of the other states in the study. A review of Ohio data indicated that from 1980 through 1991, 84 out of 88 counties in Ohio had at least one actively reporting station in each period. Of the four remaining counties, two had no station in the past twelve years and two others had only had one actively reporting station for the past two years.

Improper data aggregation and model structure are always concerns. Improper model structure refers to not including important independent variables or not using appropriate forms of the independent variables. Since the survey data and weather data are unplanned (ie., not controlled) it is often difficult to determine their real effect on yield and their appropriate model terms. The range of data values is often smaller than desired, and the survey and weather data that naturally occur in an uncontrolled setting may be correlated which reduces their usefulness (Draper and Smith 1981). This study was limited to evaluating five different model forms incorporating precipitation data and regular survey variables into the framework of the new Objective Yield multiple regression models at the regional level. Since the new OY models show improved performance using aggregated survey data values at the regional level it was anticipated that this would also prove to be an effective method for aggregating weather data.

DATA

Precipitation Data

Precipitation values used in the models represent accumulated precipitation in inches for the growing season at the regional level. For the month of August, the growing season is defined as the period from April 1 through July 31. For September, the growing season is the period from April 1 through August 31. The variable is constructed as follows:

$$P_t = \frac{\sum_{s=1}^{S} A_{ts} R_{ts}}{\sum_{s=1}^{S} A_{ts}} , \qquad (1)$$

where

P_t = the estimated accumulated precipitation over the growing season for the region, year t,

S = the number of states covered,

R_{ts} = the estimated accumulated precipitation over the growing season for year t, state s, and

A_{ts} = the Agricultural Statistics Board (ASB) acres for harvest for year t, state s.

Further:

$$R_{ts} = \frac{\sum_{d=1}^{D_s} A_{tsd} E_{tsd}}{\sum_{d=1}^{D_s} A_{tsd}} ,$$

where

A_{tsd} = the Agricultural Statistics Board (ASB) acres for harvest for year t, state s, district d,
D_s = the number of districts per state s, and

$$E_{tsd} = \frac{1}{W_{tsd}} \sum_{w=1}^{W_{tsd}} U_{tsdw} ,$$

where

E_{tsd} = the average station accumulated precipitation for year t, state s, district d,
W_{tsd} = number of weather stations for year t, state s, district d, and
U_{tsdw} = accumulated precipitation for year t, state s, district d, weather station w.

Survey Data

The construction of the survey data is discussed by Birkett (1990). For the month of August, the independent variable is the estimated number of lateral branches per eighteen square feet. For September, the independent variable is the estimated number pods with beans per eighteen square feet. The State-level estimates for August are constructed as follows:

$$F_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} B_{tsj} L_{tsj} ,$$

where

m_{ts} = the number of samples for year t, state s, for j in J, where
j = the subset of samples classified in maturity categories 2-6 (or 1-6 in the southern states), in J,
B_{tsj} = plants per 18 square feet for year t, state s, sample j,

L_{tsj} = lateral branches per plant for year t, state s, sample j, and
F_{ts} = number of lateral branches per 18 sq. feet for year t, state s.

The state level estimates are combined to the regional level with current Agricultural Statistics Board (ASB) acres harvested used as the weight as follows:

$$Z_t = \frac{\sum_{s=1}^{S} A_{ts} F_{ts}}{\sum_{s=1}^{S} A_{ts}}, \qquad (2)$$

where

A_{ts} = the ASB acres for harvest for year t, state s.

All of the definitions are the same for September except O_{tsj} is substituted for L_{tsj} and Q_{ts} is substituted for F_{ts}, where

O_{tsj} = pods with beans per plant, year t, state s, sample j, and
Q_{ts} = estimated pods with beans per 18 sq. feet year t, state s.

Only samples classified in maturity categories 6-9 are used to estimate Q_{ts}.

Board Yield

The regional Board yield values for each year which were derived from data from this set of six states is not actually published by the Agricultural Statistics Board (since this set represents a subset of the total number of states which participate in the Objective Yield Program). The Board yield values used for this analysis were calculated as follows:

$$Y_t = \frac{\sum_{s=1}^{S} A_{ts} V_{ts}}{\sum_{s=1}^{S} A_{ts}}, \qquad (3)$$

where

Y_t = final regional board yield for year t,
V_{ts} = state board yield for year t, state s.

METHODOLOGY

Statistical analysis methods used to evaluate the performance of precipitation data in combination with survey data are correlation and regression analysis. Multiple linear regression models with associated diagnostics for model fit and forecast accuracy were examined. Correlation analysis was used initially to select the optimal survey variables for use in the models. This was done by determining whether positive linear relationships exist between the dependent and independent variables. In addition to this, the correlation between independent variables was examined to detect possible collinearity problems. If high correlations exist between independent variables, then variance estimates can be large. One possible remedy is to include only the independent variable that has the highest correlation with the dependent variable in the model (Neter, Wasserman and Kutner 1983).

The following regression models were examined for each month.

Model 1*A*: $Y_t = \beta_o + \beta_1 Z_t + \epsilon_t$

Model 1*B*: $Y_t = \beta_o + \beta_1 P_t + \epsilon_t$

Model 2*A*: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \epsilon_t$

Model 2*B*: $Y_t = \beta_o + \beta_1 P_t + \beta_2 P_t^2 + \epsilon_t$

Model 3: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 P_t + \epsilon_t$

Model 4: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \epsilon_t$

Model 5: $Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \beta_4 Z_t P_t + \epsilon_t$

Model 1A is used by NASS for the August and September forecasts. Model 2A was used in September, 1991, on an experimental basis. Model 5 is the most extensive. It is a mixture model that considers response surfaces for additive and interacting independent variables. The assortment of models were examined so that comparisons in performance could be made. The method of least squares is used to estimate the parameters in the approximating polynomial (Neter, Wasserman and Kutner 1983).

Model Evaluation Criteria

The primary model evaluation criterium is a set of prediction intervals (PI) for the years 1988, 1981 and 1990. These years correspond to the minimum, median and maximum six state regional soybean yields, respectively, over the 12 years in the study. A second criterium is the adjusted coefficient of determination, R_a^2 which provides a measure of correspondence between predicted and actual yields. Both the PI and R_a^2 are based on the sum of squared differences from the least squares analysis used to derive the model parameters. Two other criteria are provided which are based on the absolute relative differences (ARD) between the predicted and actual yields (Sebaugh and Cotter 1983; House 1977). The regression models are not derived to produce minimum ARD over years, but instead are designed to minimize the sum of squared differences. Nevertheless, a minimum ARD is an important goal for prediction. These criteria evaluate which of the least squares models tend to produce the lowest ARD. Each of these evaluation criteria is further defined below.

1. The prediction interval (PI) refers to half of the 1 - α confidence interval length for the predicted value of a future Y for a given future year o. That is

$$P\,I = t(1-\frac{\alpha}{2};n-1-p)SD(\hat{Y}_o),$$

where

$$SD(\hat{Y}_o)=s[(x_o'(X_o'X_o)^{-1}x_o) + 1]^{\frac{1}{2}},$$

and

s = (residual MSE)$^{1/2}$,

x_o = relevant p-dimensional row vector of independent variables for year o (for example, in Model 3: p = 3, x_o = [1, Z_o, P_o]),

X_o = relevant (n-1 x p) matrix of independent variables (excludes x_o),

n = number of years,

p = number of parameters, and

α = the significance level.

The X_o matrix excludes the row vector x_o, so that the PI reflects the accuracy expected in an operational model where current year data are not included in the model development. A significance level of 0.32 was used for this study, which provides t values near 1.0. Consequently, the future Y will fall within the calculated PI of the predicted Y approximately 68% of the time.

2. R_a^2 is used as a goodness-of-fit test for each model with an adjustment made for the corresponding degrees of freedom (Draper and Smith 1981). R_a^2 is calculated as:

$$R_a^2 = 1 - \frac{(RSS_p)/(n - p)}{(CTSS)/(n - 1)},$$

where

RSS_p = the residual sum of squares taking the changing number of parameters jnto account,
CTSS = the corrected total sum of squares,
n = the number of years, and
p = the number of parameters.

3. The average absolute relative difference (ARD) is calculated as:

$$\overline{ARD} = \frac{1}{n}\sum_{t=1}^{n} |RD|_t$$

where

$$|RD|_t = 100\frac{|\hat{Y}_t - Y_t|}{Y_t}$$

Y_t = regional level Board Yield, year t, and

$\hat{Y}_t$ = regional level predicted yield, year t.

The predicted yield $(\hat{Y}_t)$ is based on a model that does not include data from the forecast year. This statistic is a measure of forecast reliability that provides an empirical indication of how closely the model predicted values come within Board yields on a percentage basis, without any distributional assumptions.

4. The number of years the ARD is less than 5% provides an empirical basis for comparing how consistently predicted yields are within 5% of the Board yield.

Outlier Identification

Since the purpose of the models is to make forecasts, the rstudent statistic (also called the studentized residual) was used to help identify outliers to be excluded from the model. This statistic was first recommended in Belsley, Kuh and Welsh (1980). It is similar to the standardized residual, which is defined as:

$$r_{si} = \frac{r_i}{s\sqrt{1-h_i}} \text{ ,}$$

where

r_i = i^{th} residual,
s = (residual MSE)$^{1/2}$, and
h_i = $x_i'(X'X)^{-1}x_i$.

In the rstudent statistic, s is replaced by s(i). S(i) is the estimate of σ with the i^{th} observation deleted. In a forecasting model, rstudent measures how many prediction standard errors the forecast is from the observed Y. Observations with absolute values of rstudent greater than 3.0 were identified as outliers. The rstudent statistic is distributed closely to the t-distribution with n-p-1 degrees of freedom.

The result of the examination of the rstudents found that in September only, 1980 is an outlier for Models 2A, 4, and 5. To test the improvement that occurs within each of these models, 1980 was excluded in a second regression analysis (refer to Table 4).

RESULTS

Correlation Analysis

A correlation analysis was conducted over the twelve years of data prior to performing the regression analysis to measure the degree of linearity between the dependent and independent variables and between the independent variables.

TABLE 1: AUGUST CORRELATION ANALYSIS

	PRECIP	LATERALS
YIELD	0.43	0.79
p value	0.17	0.00
PRECIP		0.33
p value		0.30

In August, the number of lateral branches per plant per eighteen square feet area has a significant positive correlation with the Board yield estimates: Pearson's R = 0.79 with a significance level of approximately 0.00. The correlation of Board yields with the precipitation for April 1 through July 31 has a Pearson's R of 0.43 and a significance level of 0.17. The survey and precipitation variables are correlated with Pearson's R = 0.33, but this correlation is not significant (p = 0.30).

TABLE 2: SEPTEMBER CORRELATION ANALYSIS

	PRECIP	PODS
YIELD	0.56	0.81
p value	0.06	0.00
PRECIP		0.42
p value		0.18

For September, the number of pods per eighteen square feet area has significant positive correlation with the Board yield estimates: Pearson's R = 0.81, with a significance level of approximately 0.00. The correlation of Board yields with total accumulated precipitation for April 1 through August 31 has a Pearson's R of 0.56 and a significance level of 0.06. The number of pods with beans per eighteen square feet and precipitation are not independent since Pearson's R = 0.42, but this correlation is not significant (p = 0.18).

Regression Analysis

The evaluation criteria statistics are presented in Table 3 for August and in Table 4 for September. Based primarily on comparisons of the PIs, the best model for August is Model 1A, which is the model currently being used by NASS to provide August forecasts. Model 2A is a close second, especially when considering the empirical ARD criteria. Model 1A consistently has the lowest prediction intervals (PI) of 2.93, 2.58 and 2.58 for years when the minimum, median and maximum yields occur (1988, 1981 and 1990) respectively. Adding the precipitation term (Model 3) increased the length of the prediction intervals by approximately ten percent for all three years, but did produce an equivalent R_a^2 and slightly lower ARD values.

TABLE 3: AUGUST REGRESSION MODEL EVALUATION

MODEL	$\hat{b}_o$	$\hat{b}_I$	PI*	R_a^2 (%)	$\overline{ARD}$	# yrs ARD < 5.0%
Model 1A: Laterals	14.76	0.3328	2.93 2.58 2.58	.63	5.42	6
Model 1B: Precip	27.84	0.4838	4.50 4.19 4.14	.18	8.18	3
Model 2A: Lats, Lat^2	-14.99	1.3335 -0.0082	3.24 2.61 2.61	.64	4.64	8
Model 2B: Precip, Precip^2	9.37	3.1785 -0.0935	4.94 4.39 4.23	.24	6.10	7
Model 3: Lats, Precip	13.16	0.2100 0.3073	3.21 2.86 2.88	.63	5.21	7
Model 4: Lats, Lat^2, Precip	-13.73	1.2851 -0.0079 0.0151	3.46 3.07 3.08	.61	4.69	8
Model 5: Lats, Lats^2, Precip, Interaction	-11.65	1.2947 -0.0086 -0.3922 0.0069	3.90 3.60 3.34	.56	4.67	7

* Note (for Tables 3 and 4) : PI is evaluated for years when yield is at minimum, median and maximum for 1988, 1981 and 1990 respectively. Outliers were identified by examining the rstudent statistics having an absolute value greater than 3.0.

TABLE 4: SEPTEMBER REGRESSION MODEL EVALUATION

MODEL	$\hat{b}_o$	$\hat{b}_I$	PI*	R_a^2	$\overline{ARD}$	# yrs
				(%)		ARD < 5.0%
Model 1A: Pods	-4.20	0.0276	2.63	.65	5.42	6
			2.50			
			2.53			
Model 1B: Precip	22.81	0.6526	4.19	.31	7.54	4
			3.93			
			3.83			
Model 2A: Pods,	-201.02	0.3020	2.46	.71	4.35	8
Pods^2		-0.0001	2.38			
			2.39			
Outlier 1980 removed:						
Model 2A: Pods,	-358.10	0.5129	1.64	.90	2.66	10
Pods^2		-0.0002	1.50			
			1.50			
Model 2B: Precip,	-3.55	3.5931	4.57	.37	7.04	7
Precip^2		-0.0797	4.10			
			3.83			
Model 3: Pods,	-4.67	0.0238	2.88	.68	5.05	5
Precip		0.3136	2.74			
			2.63			
Model 4: Pods,	-161.00	0.2435	2.82	.70	4.30	7
Pods^2,		-0.0001	2.69			
Precip		0.1900	2.59			
Outlier 1980 removed:						
Model 4: Pods,	-353.53	0.5064	1.89	.89	2.65	10
Pods^2,		-0.0002	1.77			
Precip		0.0157	1.69			
Model 5 Pods,	-215.61	0.2613	3.16	.71	3.95	7
Pods^2,		-0.0001	2.74			
Precip,		4.8977	2.80			
Interaction		-0.0034				
Outlier 1980 removed:						
Model 5 Pods,	-376.00	0.5021	2.11	.89	2.24	11
Pods^2,		-0.0001	1.82			
Precip,		2.9245	1.87			
Interaction		-0.0021				

In September, Model 2A: Pods and Pods^2, the quadratic model, is the best model when evaluated in terms of prediction intervals. It is the simplest and most cost efficient model. It has relatively low prediction intervals of 2.46, 2.38 and 2.39 for 1988, 1981 and 1990 respectively; a relatively high R_a^2 of .71; a relatively low average ARD value of 4.35 (on average less than 5.0); and predicts within 5% of Board yield eight out of twelve years. Model 4, which adds the precipitation term investigated for this study to the quadratic model, has values comparable to those of Model 2A for R_a^2, average ARD, and number of years the ARD is less than five percent. But, the prediction intervals for Model 4 are approximately ten percent larger that those for Model 2A.

A further check was made to see whether any noticeable improvement would occur within models that produced an outlier if that outlier was removed. In September only, Models 2A, 4 and 5 showed that 1980 is an outlier. After that observation was removed, Model 2A showed extremely good results. The prediction intervals (PI) are 1.64, 1.50, and 1.50 for the evaluation at the same arbitrary points (years 1988, 1981 and 1990) respectively. It has a very high R_a^2 value of .90, an extremely low average ARD value of 2.66% and predicts within 5% of the Board Yield ten years out of eleven years. Model 4, which included precipitation, performs almost as well as Model 2A in terms of R_a^2 (.89), average ARD (2.65), and number of years ARD < 5% (10). Model 2A, however, consistently has smaller prediction intervals.

See Appendix A for descriptive (summary) statistics of the dependent and independent variables and Appendix B for a graphic representation of the data.

CONCLUSIONS

In August, there is no evidence that a change from the univariate survey data model is warranted. In September, the quadratic model, using pods and pods^2 (2A), shows definite improvement in all evaluation criteria over the univariate model (1A). Adding the precipitation term investigated for this study to the quadratic model provides values of R_a^2, average ARD, and number of years the ARD is less than five percent which are comparable to the quadratic model values. But, the quadratic model consistently has smaller prediction intervals.

RECOMMENDATIONS

This analysis has shown that soybean yield forecasts were not improved using this particular precipitation time frame. In August, the univariate model containing survey data is sufficient for estimating yield. In September, the quadratic survey data model is preferred.

The following recommendations are made.

1. Investigate other precipitation time frame terms such as monthly total accumulated precipitation for May, June, July and August as well as row spacing and planting dates.

2. Analyze other crops such as corn, cotton and wheat to determine if weather data can improve forecast accuracy at the regional and state levels.

3. Continue research to analyze the feasibility of a real time weather data system whereby preliminary precipitation data are used in conjunction with historical weather data to make current yield forecasts.

BIBLIOGRAPHY

Belsley, David A, Kuh, Edwin, Welsh, R.E., (1980), Regression Diagnostics, John Wiley & Sons.

Birkett, Thomas R., (1990) "The New Objective Yield Models for Corn and Soybeans", SMB Staff Report Number SMB-90-02, U.S. Department of Agriculture.

Draper, N.R., Smith, H., (1981), Applied Regression Analysis, John Wiley & Sons, Second Edition.

House, Carol C., (1977) "A Within-Year Growth Model Approach to Forecasting Corn Yields", Crop Reporting Board, Economics, Statistics, and Cooperatives Service, U.S. Department of Agriculture.

Kaiser, Mark, Sebaugh, Jeanne L., (1984) "Methods for the Evaluation of Real-Time Weather Data for use in Crop Yield Models: An Application to North Dakota", SRD Report Number AGES840424, U.S. Department of Agriculture.

Kestle, Richard A., (1981) "Analysis of Crop Yield Trends and Development of Simple Corn and Soybean "Straw Man" models for Indiana, Illinois, and Iowa. AgRISTARS Yield Model Development Project. Document YMD-2-11-1 (80-11.1), ESS Staff Report AGES810114, U.S. Department of Agriculture.

Maas, Stephan J., (1982) "Forecasting Yields Using Weather-Related Indices", SRD Staff Report Number YRB 8-2-08, U.S. Department of Agriculture.

National Oceanic and Atmospheric Administration, (1987) "TD-3200 Summary of Day Co-operative", U.S. Department of Commerce.

Neter, John, Wasserman, William, Kutner, Michael H., (1983), Applied Linear Regression Models, Richard D. Irwin, Inc.

Sanderson, Fred H., (1942) "Use of Condition Reports and Weather Data in Forecasting the Yield per Acre of Wheat", SMB Staff Report Number YRB 42-01, U.S. Department of Agriculture.

Searle, S.R., (1971) Linear Models, John Wiley & Sons.

Sebaugh, Jeanne L., (1981) "Evaluation of "Straw Man" Model 1, the Simple Linear Model, For Soybean Yields in Iowa, Illinois and Indiana", SRD Staff Report Number AGES811214, U.S. Department of Agriculture.

Sebaugh, Jeanne L., (1981) "One, Two and Three Line Segment "Straw Man Models, Soybean Yields in Iowa, Illinois and Indiana", SRD ESS Staff Report Number AGESS810514, U.S. Department of Agriculture.

Sebaugh, Jeanne L., Cotter, James J., (1983) "Comparison of the CEAS and Thompson-type Models for Soybeans Yields in Iowa, Illinois and Indiana", SRS Staff Report Number AGES830613, U.S. Department of Agriculture.

Sebaugh, Jeanne L., (1983) "Evaluation of the Feyerherm '81 Spring Wheat Models for Estimating Yields in North Dakota and Minnesota", SRS Staff Report Number AGES830609, U.S. Department of Agriculture.

Warren, Fred B., (1990) "An Operational Test Using Weather Data to Forecast Corn Ear Weight, 1988", SRB Staff Report Number SRB-90-05, U.S. Department of Agriculture.

APPENDIX A

AUGUST DESCRIPTIVE STATISTICS

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
YIELD	12	35.12	3.71	421.46	27.82	38.72
PRECIP	12	15.04	3.26	180.54	8.09	20.04
LATS	12	61.19	8.86	734.24	44.56	76.96

SEPTEMBER DESCRIPTIVE STATISTICS

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
YIELD	12	35.12	3.71	421.46	27.82	38.72
PRECIP	12	18.86	3.17	226.32	11.77	24.35
PODS	12	1424.00	108.38	17087.00	1279.00	1605.00

APPENDIX B

Graphic Representation of the Data

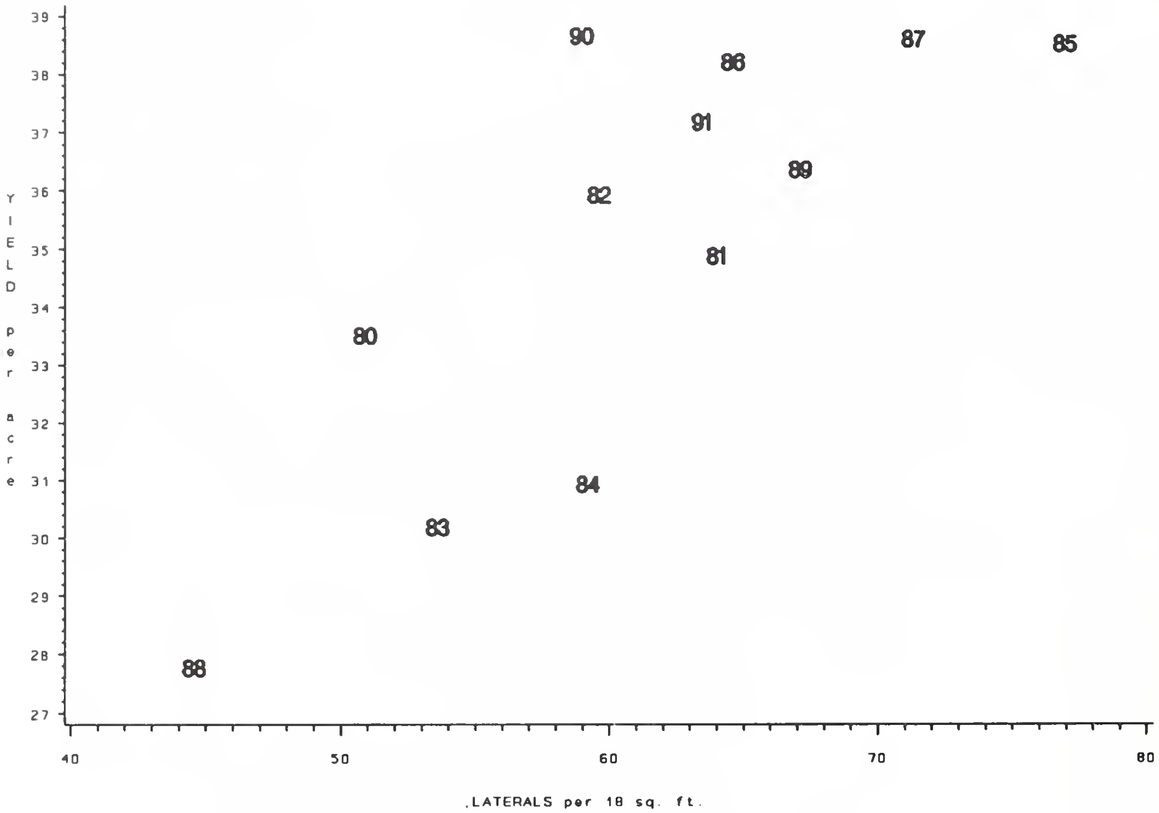
Initially, the data were aggregated to the regional level. There are twelve observations (one for each year in this analysis) per variable. Plots of the independent variable versus the dependent variable and versus each other for each month are shown in Graphs 1 to 6 on the following pages.

These plots were reviewed to determine whether or not there are any visible trends in the data which could have had a bearing on model structure. Generally for both August and September, the relationship between survey data and yield is positive and linear. The higher the count of lateral branches per eighteen square feet and the higher the count of pods per eighteen square feet, the higher the estimated yield.

The relationship is more difficult to interpret for precipitation data. In the August data plot of precipitation versus yield (Graph 3) precipitation ranging from thirteen to twenty inches for the growing period of April 1 through July 31 could produce the same level of about a thirty-seven bushel per acre regional yield estimate. In September (Graph 4), a level of about eighteen inches of precipitation for the growing period of April 1 through August 31 could produce a range of thirty to thirty-nine bushels per acre as a regional yield estimate.

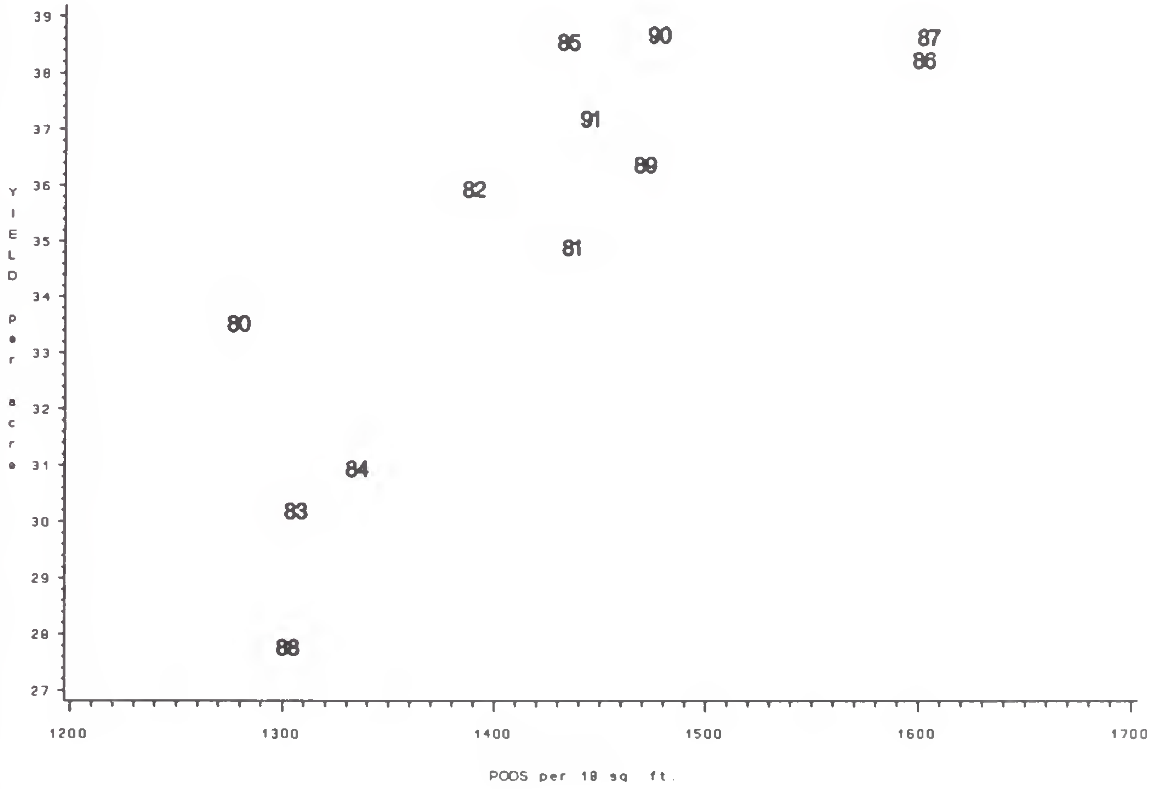
In the plots of the independent variables against each other there is no clear pattern that could suggest a linear relationship between precipitation and survey data.

GRAPH 1: AUGUST DATA PLOT OF LATERAL BRANCHES per 18 sq. ft vs. YIELD



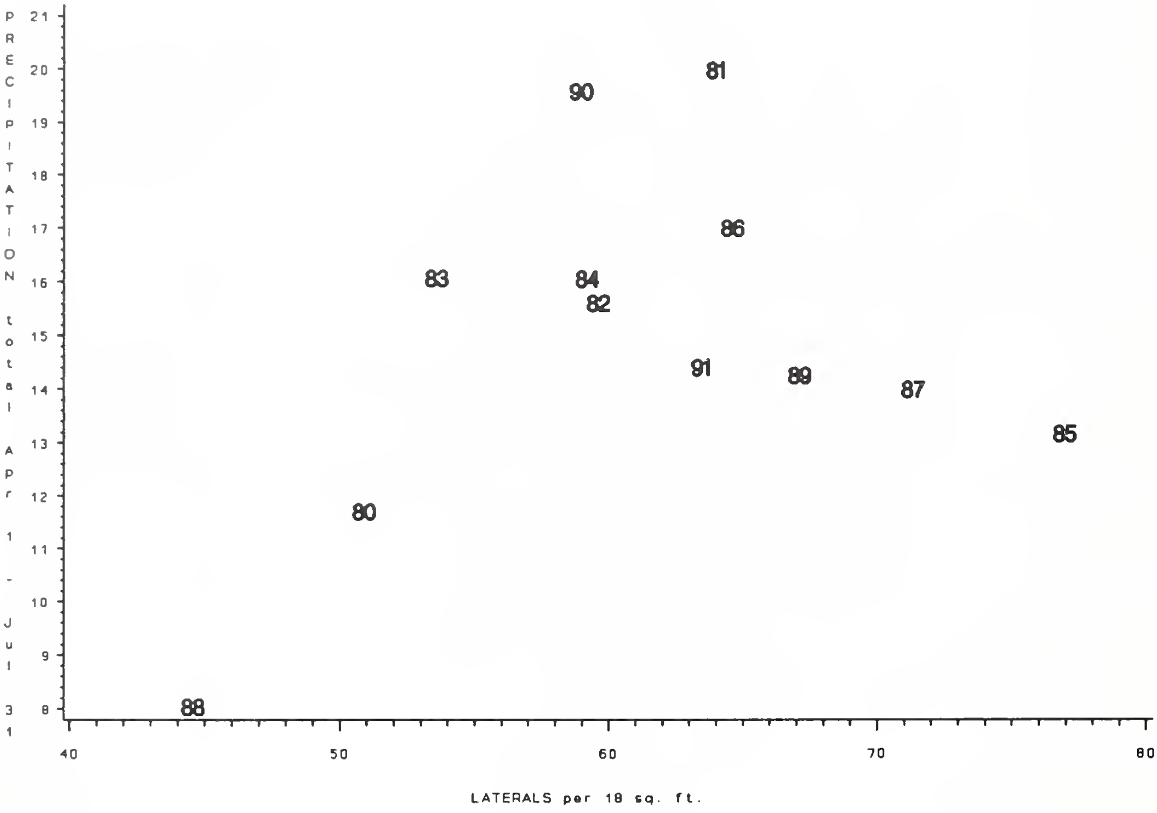
Numbers plotted represent year of occurrence

GRAPH 2: SEPTEMBER DATA PLOT OF PODS per 18 sq. ft vs. YIELD



Numbers plotted represent year of occurrence

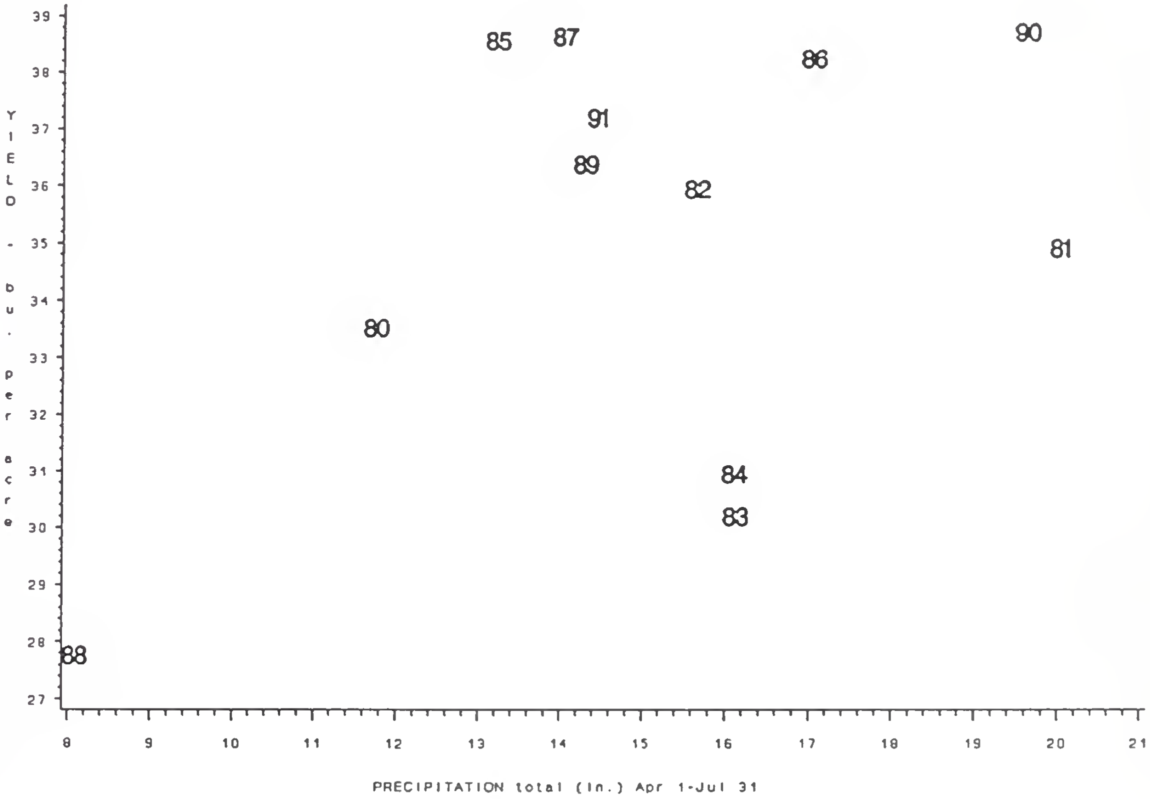
GRAPH 3: AUGUST PLOT OF THE DATA PRECIPITATION vs. YIELD



Numbers plotted represent year of occurrence

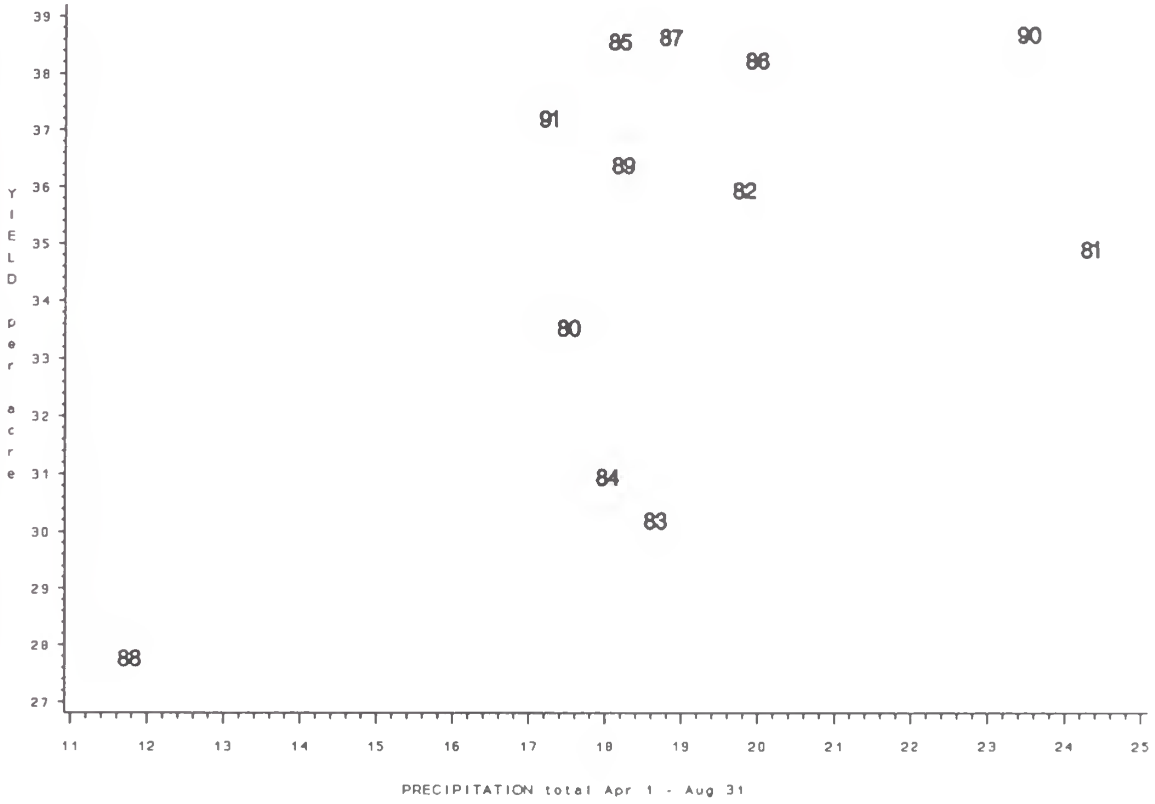
Contour

GRAPH 3: AUGUST PLOT OF THE DATA PRECIPITATION vs. YIELD



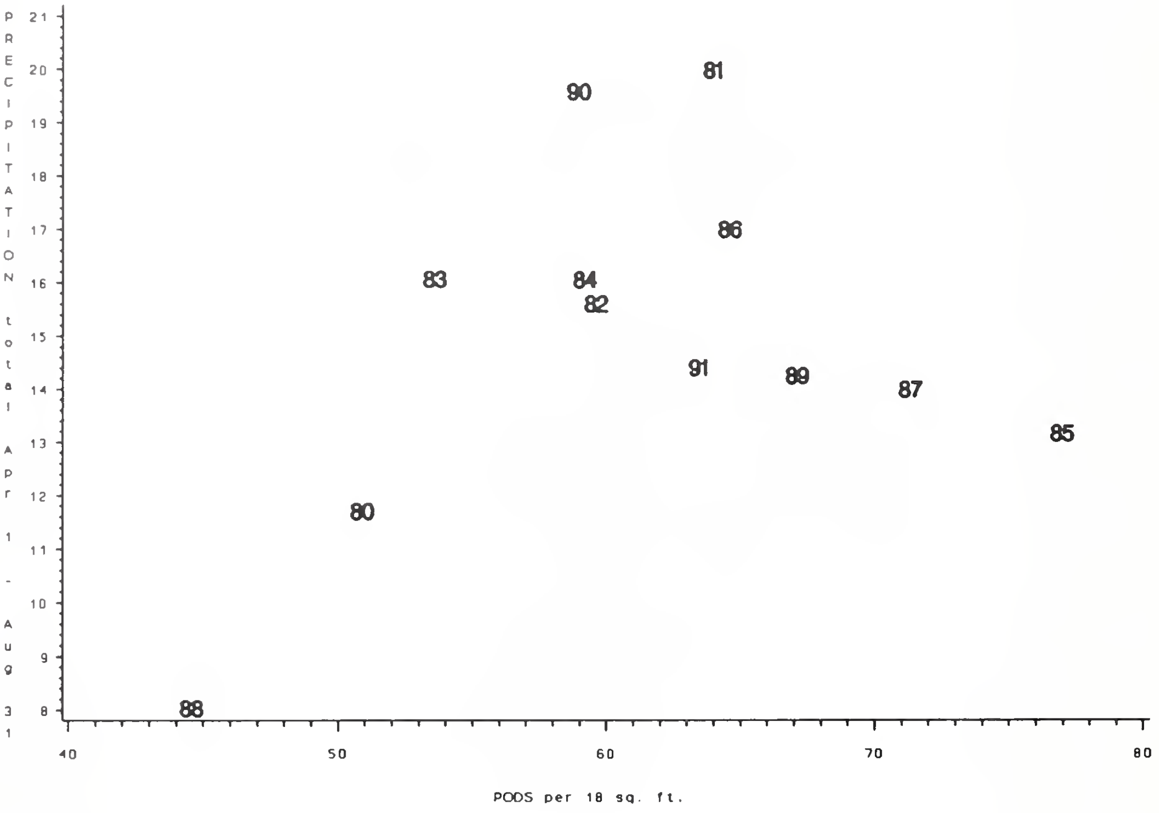
Numbers plotted represent year of occurrence

GRAPH 4: SEPTEMBER PLOT OF THE DATA PRECIPITATION vs. YIELD



Numbers plotted represent year of occurrence

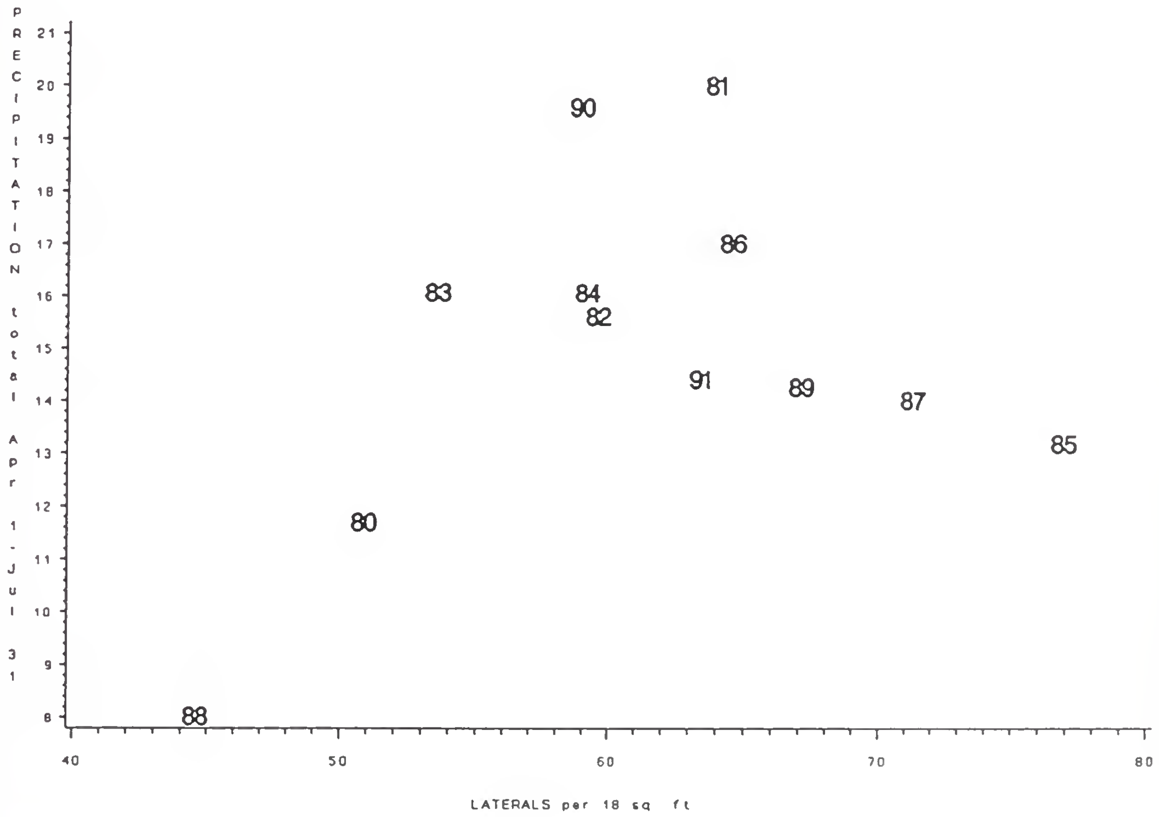
GRAPH 5: AUGUST PLOT OF THE DATA LATERALS vs. PRECIPITATION



Numbers plotted represent year of occurrence

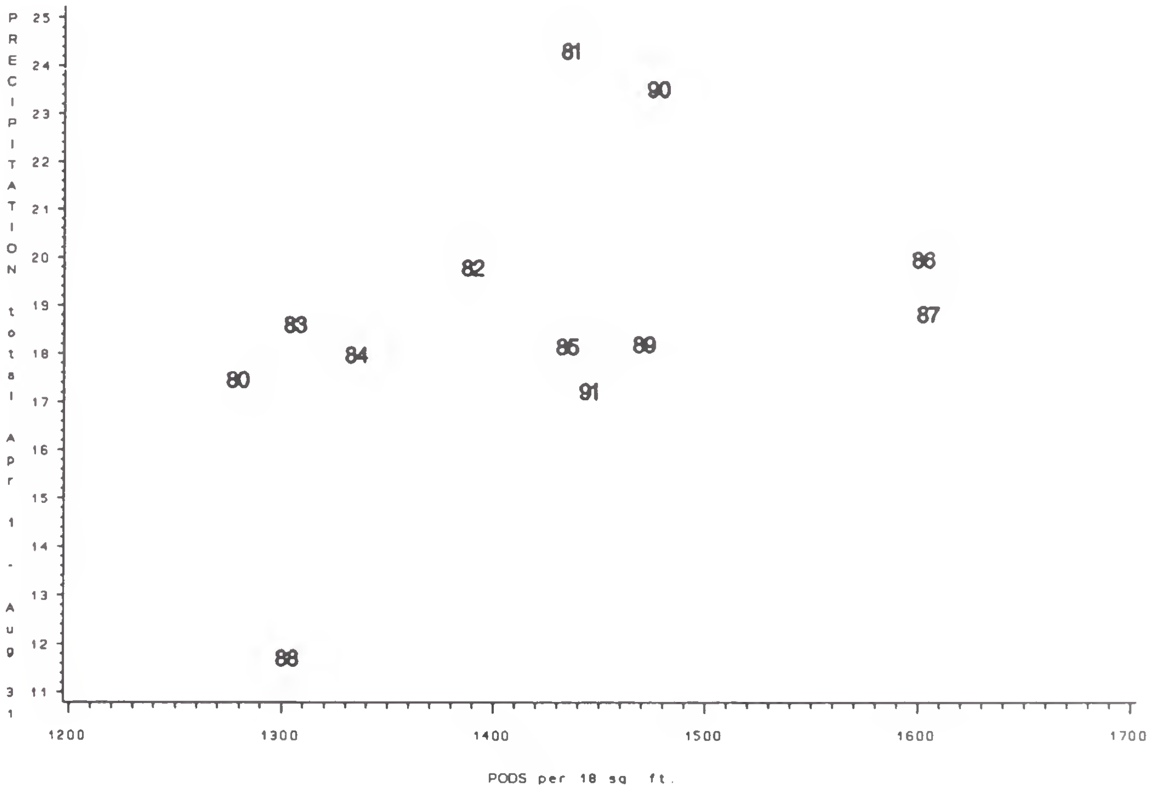
[illegible]

GRAPH 5: AUGUST PLOT OF THE DATA LATERALS vs. PRECIPITATION



Numbers plotted represent year of occurrence

GRAPH 6: SEPTEMBER PLOT OF THE DATA PODS vs. PRECIPITATION



Numbers plotted represent year of occurrence

☆U.S. Government Printing Office : 1992 - 341-378/60949